第8章　キャプション生成から文書整理，QAボットまで…

視覚言語モデルVLMで画像からテキストを生成してみる

石垣 達也

表1　視覚言語モデル（VLM）の応用範囲

タスク	例
画像キャプション生成	商品の写真やSNS投稿画像に自動で説明文を付ける
質問応答（VQA）	工場の機械写真を見て「安全装置は作動しているか？」と尋ねる
文書整理	紙資料をスキャンして，どの種類の文書かを自動仕分け
会話支援	チャットボットに画像を送って「これはどう扱えばよい？」と聞く

ここ数年，AIの世界では，視覚と言語を同時に理解できるモデルが注目を集めています．これは視覚言語モデル（Vision-Language Models：VLM）と呼ばれています．従来，画像認識はコンピュータ・ビジョンの領域，文章生成は自然言語処理の領域とそれぞれの領域に分かれていました．しかしVLMはこの2つを橋渡しし，画像や動画を理解して自然な言葉で説明したり，質問に答えたりできます．

本稿では3つのモデルの特徴を追いながら，実務で役立つコード例も紹介します．

VLMの使いどころ

VLMでは，従来の言語モデルのようにテキストを入力として与えるだけでなく画像も入力として与え，画像を理解しながらさまざまなタスクを解く点が特徴です．応用範囲は非常に広く，**表1**のようなタスクが想定されています．

①CLIP…テキストと画像を扱うLLMのブレークスルー

OpenAIが2021年に公開したCLIPは，VLMのブレークスルーとなりました．CLIPは画像とテキストを同じベクトル空間にマッピングするという発想で，400万組以上の画像–テキスト・ペアを学習しています．

CLIPの魅力は，Zero-shot学習でさまざまな分類ができる点です．Zero-shot学習とは，オブジェクトや概念を認識して分類するようにトレーニングされていて，対象物を事前に学習していなくても，追加学習しなくても分類できる学習方法です(1)．例えば「これは犬ですか猫ですか？」と分類したい場合，犬や猫のラベルをテキスト・エンコーダに入力して埋め込みを得れば，画像の埋め込みとの類似度を比較するだけで答えられます．追加学習は不要です．

● 動作イメージ

CLIPは事前学習と推論という2つのステップで構成されています．

▶事前学習

大量のテキストと画像のペア（データセット）を用意し，学習を行うというステップを経ます．

データセット作成では，例えば，犬の画像に対して，「犬の画像」というテキストを用意します．このようなペアはウェブ上にあるデータから大量のペアを作成することができます．画像とテキストはそれぞれ画像エンコーダとテキスト・エンコーダという行列演算に通され固定長ベクトルに変換されます．**図1**において，例えばT_1，T_2は用意した1つ目と2つ目のテキストに対応する固定長ベクトルです．同様にI_1，I_2も画像に対応する固定長ベクトルです．

学習段階では，テキスト・エンコーダと画像エンコーダを同時に訓練します．基本的な考え方は意味が対応する画像とテキストのベクトルは近づけ，意味が異なる組み合わせのベクトルは離すというものです．例えば，犬の画像と「犬の画像」というテキストは近い位置にマッピングされるように，逆に犬の画像と「猫の画像」というテキストは離れた位置にマッピングされるように学習します．この仕組みを実現するために，対照学習（contrastive learning）と呼ばれる技術(2)が用いられています．

▶推論

CLIPの推論では画像に写っているものを分類するタスクを解くことができます．例えば，画像を与えると「飛行機」「車」「犬」…「鳥」のラベルのうちいずれかを選ぶという設定が考えられます．